

# (19) United States

# (12) Patent Application Publication (10) Pub. No.: US 2016/0293167 A1 Chen et al.

### Oct. 6, 2016 (43) **Pub. Date:**

#### (54) SPEAKER RECOGNITION USING NEURAL **NETWORKS**

(71) Applicant: Google Inc., Mountain View, CA (US)

(72) Inventors: Yu-hsin Joyce Chen, Mountain View, CA (US); Ignacio Lopez Moreno, New York, NY (US); Tara N. Sainath, Jersey City, NJ (US); Maria Carolina Parada San Martin, Palo Alto, CA

(21) Appl. No.: 15/179,717

(22) Filed: Jun. 10, 2016

#### Related U.S. Application Data

- Continuation-in-part of application No. 14/224,469, filed on Mar. 25, 2014, now abandoned.
- Provisional application No. 62/174,799, filed on Jun. 12, 2015, provisional application No. 61/889,359, filed on Oct. 10, 2013.

#### **Publication Classification**

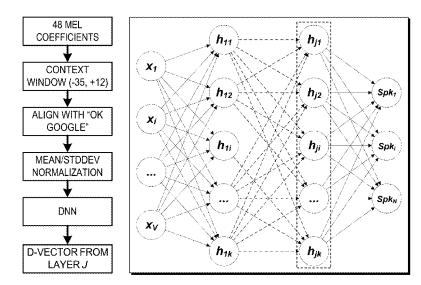
(51) Int. Cl. G10L 17/18 (2006.01)G10L 17/22 (2006.01)G10L 17/08 (2006.01)

U.S. Cl. CPC ...... G10L 17/18 (2013.01); G10L 17/08 (2013.01); G10L 17/22 (2013.01)

(57)ABSTRACT

Methods, systems, and apparatus, including computer programs encoded on a computer storage medium, for performing speaker verification. In one aspect, a method includes accessing a neural network having an input layer that provides inputs to a first hidden layer whose nodes are respectively connected to only a proper subset of the inputs from the input layer. Speech data that corresponds to a particular utterance may be provided as input to the input layer of the neural network. A representation of activations that occur in response to the speech data at a particular layer of the neural network that was configured as a hidden layer during training of the neural network may be generated. A determination of whether the particular utterance was likely spoken by a particular speaker may be made based at least on the generated representation. An indication of whether the particular utterance was likely spoken by the particular speaker may be provided.





## **Topology**

- c = 48 mel-filterbanks
- l = 35, r = 12 contextframes
- v = 2304 visible units
- M = 4 hidden layers
- k = 256 hidden units
- N = 3200 output speakers
- w = 787k model weights (excluding output layers)
- Rectified Linear Units
- Softmax output Layer

#### **Training**

Stochastic Gradient Descent